

Combining Group-Based Trajectory Modeling and Propensity Score Matching for Causal Inferences in Nonexperimental Longitudinal Data

Amelia Haviland
Rand Corporation

Daniel S. Nagin
Carnegie Mellon University

Paul R. Rosenbaum
University of Pennsylvania

Richard E. Tremblay
International Laboratory for Child and Adolescent Mental
Health Development, INSERM U669, University College
Dublin, and University of Montréal

A central theme of research on human development and psychopathology is whether a therapeutic intervention or a turning-point event, such as a family break-up, alters the trajectory of the behavior under study. This article describes and applies a method for using observational longitudinal data to make more transparent causal inferences about the impact of such events on developmental trajectories. The method combines 2 distinct lines of research: work on the use of finite mixture modeling to analyze developmental trajectories and work on propensity score matching. The propensity scores are used to balance observed covariates and the trajectory groups are used to control pretreatment measures of response. The trajectory groups also aid in characterizing classes of subjects for which no good matches are available. The approach is demonstrated with an analysis of the impact of gang membership on violent delinquency based on data from a large longitudinal study conducted in Montréal, Canada.

Keywords: causal inference, trajectories, propensity scores

A central theme of research on human development and psychopathology is whether a therapeutic intervention or a turning-point event, such as a family break-up, alters the trajectory of the behavior under study. As it is often either unethical or impractical to pursue research on this theme through experimental methods, researchers collect rich nonexperimental longitudinal data and implement methods appropriate for causal analysis of observational data. Although there are always threats to the validity of giving causal interpretation to results from observational data because of selection into the “treatment” group, the threats take a particular form in developmental research based on observational longitudinal data. The problem arises when treatment affects the

future direction of the behavior under study, and in addition prior pathways of the behavior predict both entry into treatment and the future direction of the behavior. For instance, early antisocial behavior predicts both later antisocial behavior and school failure, whereas school failure also can affect later antisocial behavior (Maguin, Loeber, & LeMahieu, 1993; Nagin, Pagani, Tremblay, & Vitaro, 2003; Pagani, Boulerice, Vitaro, & Tremblay, 1999).

This article describes and applies a method for using observational longitudinal data to make more transparent causal inferences about the impact of a therapeutic intervention or a turning-point event on developmental trajectories of behavior. This transparency is achieved by modeling our approach after some key attributes of experiments. Although the developmental research scenario we describe has particular goals and challenges, it also has particular benefits. The method we propose, which combines group-based trajectory modeling with propensity score matching, is designed to answer the types of research questions just described and make use of the strengths of the data. It does this in four ways. Propensity score matching is employed to create a control group that is comparable to the treated group with respect to the observed covariates. The group-based trajectory model allows us to take a developmental view of “comparable” by matching treated individuals with individuals who were not treated but who appeared to be on a similar developmental pathway for the behavior under study prior to treatment. In addition, the richness typical of data collected under these circumstances means that participants who are comparable on observed covariates are comparable on an important group of relevant covariates. Second, in addition to the propensity score matching providing more transparent causal inferences, the group-based trajectory groups provide a means of identifying

Amelia Haviland, Rand Corporation, Pittsburgh, PA; Daniel S. Nagin, Heinz School of Public Policy and Management, Carnegie Mellon University; Paul R. Rosenbaum, Department of Statistics, University of Pennsylvania; and Richard E. Tremblay, International Laboratory for Child and Adolescent Mental Health Development, INSERM U669, Paris, France; University College Dublin, Ireland; University of Montréal, Canada.

This research has been supported by National Science Foundation Grants SES-99113700 and SES-0647576, and National Institute of Mental Health Grant RO1 MH65611-01A2. It has also made heavy use of data collected with support from the Fonds du Québec pour la Recherche sur la Société et la Culture and the Fonds de Formation des Chercheurs et d'Aide à la Recherche in Quebec, Canada; the Canadian Institute for Health Research and the Social Sciences and Humanities Research Council of Canada; and the Molson Foundation.

Correspondence concerning this article should be addressed to Daniel S. Nagin, Heinz School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: dn03@andrew.cmu.edu

developmentally meaningful subgroups in the population for whom treatment effects may vary. Identifying these heterogeneous treatment effects is often an important goal or subgoal of developmental research. Third, the method we propose “keeps time in order” by dealing appropriately with the longitudinal data on both treatment involvement and the behavior of interest after the initial treatment by recognizing that they are both outcomes of initial treatment. Fourth, because the adjustment for observed covariates is transparent, it encourages and frames consideration and discussion of the possible biases from unobserved covariates that were not controlled by matching.

The approach described and applied here has been developed in detail in Haviland, Nagin, and Rosenbaum (2007), which in turn builds on Haviland and Nagin (2005). Here we provide an overview of the approach and provide an illustration of its application with a simple version of the case study presented in Haviland et al. The case study investigates whether joining a gang at age 14, the turning-point event or treatment, has a violence facilitation effect on violent delinquency at age 14 and beyond. This case study was chosen for two reasons: First, it addresses an important developmental question concerning the effect of peer influence in adolescence on the development of violence; second, criminologists have long recognized that estimates of the violence facilitation effect of gang membership are highly susceptible to confounding because those who join gangs are often the most prone to violent delinquency prior to joining the gang. These attributes exemplify the context for which the methods illustrated here were developed.

We begin with a discussion of the attributes of an experiment that we bring in to our approach and compare the proposed method with current alternatives. This discussion is followed by descriptions of the method itself and the data for the case study. We then illustrate the three stages of the method by applying it to the case study in which we investigate the effect of gang membership on violence.

Some Key Attributes of Experiments

Treatment Control

One key attribute of experiments is that treatment is under experimental control. One important aspect of control is the timing of initial treatment application. In an experiment, there is a sharp distinction between, on the one hand, baseline or pretreatment covariates that are not affected by the treatment and, on the other hand, posttreatment outcomes or responses that may be affected by the treatment (Rosenbaum, 1984). In many longitudinal observational studies, turning-point events can occur intermittently and be sustained for different amounts of time. We frame the nonexperimental study of such events in terms of an analogous experiment by carefully defining the treatment of interest, including when it starts, and consequently specifying which observations can be defined as baseline and which observations are measured after the initiation of treatment. We pay particular attention to setting the initiation of treatment at a particular time, because treatment may affect not only future development but also future exposure to the treatment. Thus, part of the effect of a treatment may be to either encourage or discourage its continued use—which, in turn, may be an important determinant of the ultimate effect of initial treatment status on development. The experimental framework we illustrate emphasizes these issues.

In the current illustration we adopt a simple approach, focusing on boys who had no experience in gangs prior to the age of 14, some of whom joined gangs at age 14. We match the boys who joined a gang at age 14 to controls who did not join a gang at age 14 and follow them forward to see what happens next. A slightly more complex approach called *risk-set matching* would study boys who joined gangs at various ages. In this alternative approach, whenever a boy joined a gang for the first time, whether at age 14 or at some other age, that boy would be paired with a control who had not yet joined a gang but who was otherwise similar up to that moment; see Li, Propert, and Rosenbaum (2001) and Lu (2005) for discussion of risk-set matching.

Good Baseline Measurements and Covariate Balance

Another key feature of a well-designed randomized experiment is comparability of the treated and controls on pretreatment covariates due to randomization of treatment status, which leads to an unbiased treatment effect estimate. The term *covariate balance* specifically refers to treated and control groups that appear comparable as groups prior to the start of treatment, here prior to age 14. In nonrandomized or observational studies, comparability may and should be assessed for observed covariates; but there is no basis for believing the groups are comparable with respect to covariates that were not measured, as there would be with a randomized experiment. Possible biases from unmeasured covariates are a central concern in observational studies and are judged by other means, such as sensitivity analysis, as illustrated in Haviland et al. (2007) and as discussed in greater detail in Rosenbaum (2002, 2005a, 2005b, 2005c).

The amalgam of propensity score matching and trajectory modeling is designed to re-create from observational data, to the limited extent possible, this key feature of a randomized experiment. This is done by attempting to take maximum advantage of two characteristics of modern longitudinal data sets: (a) the very rich set of measurements that profile the psychosocial characteristics of study participants and (b) the extended pretreatment measurement of the outcome variable. Specifically, we pair a gang joiner at age 14 to a control with a similar pattern of violent behavior prior to age 14 and with a similar probability of joining a gang, based on observed covariates measured prior to age 14. Trajectory groups summarize, in a few categories, the pattern of past violent behavior. Propensity scores estimate the chance that each boy will join a gang, summarizing many covariates in a single constructed covariate.

Our point of departure in describing the amalgam of propensity score matching and trajectory modeling laid out in Haviland et al. (2007) is immediately prior work by Haviland and Nagin (2005). Haviland and Nagin (2005) used group-based trajectory modeling alone to bring the above attribute of an experiment to the estimation of treatment effects from observational longitudinal data. Group-based trajectory modeling is designed to identify groups of individuals following approximately the same developmental trajectory over a specified period of time (e.g., ages 11 to 13) for the outcome of interest (e.g., violent delinquency).

The trajectory groups can be thought of as latent strata representing the baseline history of the outcome variable. In many contexts, the pretreatment history of the outcome variable—here, the course of violent behavior before age 14—comprises the most

important pretreatment variables predicting the posttreatment course of the outcome, and trajectory groups focus on these key covariates.

In its simplest form the analysis involves assigning individuals to the trajectory group they most likely belong to, on the basis of estimates of each individual's posterior probability of group membership. In the context of the illustrative analysis this stratifies the sample according to their developmental history of violence up to age 14. This categorization is also a key step in the analyses we present here. Gang joiners are compared to controls in the same trajectory group—that is, controls whose past violent behavior is similar to that of the joiner. This restriction is designed to reduce confounding on these key covariates, pretreatment measures of the outcome variable. The restriction may also balance additional covariates that are strongly associated with pretreatment values of the outcome.

Although use of trajectory groups alone may or may not balance covariates other than the pretreatment history of the outcome (Nieuwebeerta, Nagin, & Blokland, 2007), propensity scores were designed to achieve approximate balance on many covariates simultaneously. The combined use of trajectory groups and propensity scores is illustrated in detail in Haviland et al. (2007).

In the simplest randomized experiment, subjects are assigned to treatment or control by the independent tosses of a fair coin so that every individual has the same chance, namely 50%, of receiving treatment rather than control. In contrast, the defining feature of an observational study is that randomization is not used to assign treatments, so that some individuals are more likely to receive the treatment than others. For instance, the boys who joined gangs at age 14 did not do so at random, with equal probabilities; in fact, the boys who joined gangs at age 14 tended to be quite different from those who did not, even several years prior to age 14.

The propensity score is the conditional probability of receiving the treatment rather than the control given the observed covariates (Rosenbaum & Rubin, 1983). In the current context, the propensity score is the conditional probability of joining a gang at age 14, given the observed covariates, namely, violence prior to age 14, peer-rated popularity, mother's age at the birth of her first child, and so on. If two boys have the same propensity score given observed covariates, say a 20% chance of joining a gang at 14, then these observed covariates will be of no further use in predicting which of these two boys will join a gang at 14; so for these two boys, there will be no systematic tendency for the observed covariates to be different for the joiner and the nonjoiner. Typically, a matching algorithm insists that pairs be close on the estimated propensity score; once this is achieved, the algorithm seeks pairs that are close in other ways as well. A nontechnical survey of methods and results about propensity scores is given by Joffe and Rosenbaum (1999); for several case studies, see Dehejia and Wahba (1999), Rosenbaum, Ross, and Silber (2007), Rosenbaum and Rubin (1984, 1985), and Smith (1997). Because matching contrasts pairs of actual individuals, it is straightforward to integrate quantitative research using matching with qualitative or ethnographic research that constructs a narrative description of these same individuals (Rosenbaum & Silber, 2001). An ethnographic investigation of a limited number of matched pairs may provide insight into possible biases from covariates that were not measured and hence not controlled by the matching.

The integration of group-based trajectory modeling and propensity scores is composed of a three-stage analysis. The first stage involves estimating a group-based trajectory model for the outcome and subjects of interest. In the context of our demonstration analysis, this step involves estimation of a trajectory model of violent delinquency from ages 11 to 13 for individuals with no gang involvement over this period. In the second stage, each treated individual is matched with one or more untreated individuals. The matching of gang joiners with gang nonjoiners, carried out within trajectory group, attempts to find nonjoiners who are close on an estimate of the propensity score and on specific covariates judged to be especially important. We then check the degree of success of the matching strategy in achieving balance between the first-time gang members—the treated—and their matched counterparts who did not join gangs—the controls. In the third stage of the analysis the treatment effect of the event of interest—gang membership, in our case—is analyzed. Specifically, we examine the effect of first-time gang membership at age 14 on violence at age 14 and beyond, within and across trajectory groups.

Stratification and Heterogeneous Treatment Effects

In a randomized experiment at the experimental design stage, segments of the population of interest that are hypothesized to experience differential treatment effects can be separated into design strata. Subjects from each stratum are independently sampled and randomized into treatment. This design feature allows for straightforward calculation of strata-specific treatment effects and comparison of treatment effects across strata.

The use of trajectory groups as a basis for inference leads to the estimation of trajectory-group-specific treatment effects. This is scientifically important because a key premise of life course theories of development is that the magnitude, including the sign, of treatment effects may depend upon a person's developmental trajectory (Elder, 1985, 1998; Thornberry, Krohn, Lizotte, Smith, & Tobin, 2003). Thus, in addition to contributing to covariate balance, the trajectory group framework allows for examination of whether there are differences in treatment effects across substantively interesting groups that are differentiated by their developmental history.

Comparison With Other Methodology for Longitudinal Data

A common mistake in studying people over time is to designate certain variables as predictors and others as outcomes, where the designation is based on research objectives rather than on what causes what. More specifically, the objective of ascertaining the effect of A on B is often recast as a statistical investigation in which A is treated as a predictor of the outcome B. This specious start can become buried amid complex methodology that does nothing to remove the problem but does quite a bit to obscure it. A famous and illuminating example was carefully described by Mitchell Gail (1972) in his reassessment of two early studies that claimed that heart transplants dramatically prolonged life. The aim of these studies was to ascertain the effect of transplants (A) on the outcome, longevity (B). These early studies compared a treated group to a control group; specifically, they compared the survival

of patients who had received a heart transplant to patients who had not. These very first studies were not randomized clinical trials—unlike new drugs, surgical innovations are often tried informally before a clinical trial is mounted. Of course, the interest is in the effects of heart transplant on survival. Given this intent, it is easy to slip into the presumption that any statistical association reflects the effect of heart transplant (cause) on the outcome (survival). As it turns out, survival was much better among those who had a heart transplant. Gail raised doubts about whether this was an effect of heart transplants on survival; specifically, he argued it might have been an effect of survival on heart transplantation. Patients were entered into the studies, but they did not receive an immediate heart transplant, because they had to wait until a suitable heart became available. Indeed, if a heart became available, a transplant was performed; so patients entered the “treated group” by surviving until a heart became available, and they entered the “control group” by dying before a heart became available. Dying early prevented them from receiving a heart, and dying late permitted them to receive a heart. Gail was raising the possibility that the investigators had merely confirmed a familiar fact: Dying dramatically shortens your life expectancy. Gail then asked: How would an experiment have been done? Patients who were candidates for the same heart might be paired, and when a heart became available, one of them would receive it based on a coin flip, measuring survival from that time; then transplants could affect survival, but survival could not affect transplantation.

In the current context, any method of analysis that compares a treated group of “boys who joined gangs from ages 14 to 17” to a control group of “boys who did not join gangs from ages 14 to 17” as a cause of violence outcomes runs the risk of a parallel mistake. That is, one cannot designate gang membership as a predictor and violence as an outcome; designation does not make it so. It is for this reason that we compare 1 boy at age 14 who joined a gang at age 14 to 2 boys who were similar before age 14 but did not join a gang at age 14, without regard to later gang membership. For brevity, we later refer to this issue as “keeping time in order”; technically, it is risk-set matching (Li et al., 2001; Lu, 2005).

The error just described is facilitated by a model or software that frames longitudinal analysis as identifying aspects of the data with certain aspects of the model; for instance, the first step in running a regression is to designate certain variables as predictors and another variable as an outcome. There is no aspect of the data in the heart transplant example that would reveal the problem Gail (1972) detected; rather, it is the contrast of the manner in which the data were collected with the manner in which experimental data would have been collected that reveals the problem. Framing longitudinal data analysis in terms of an analogous, simple experiment, as we do, calls attention to the ways that a nonexperimental study falls short of a true experiment. Having limitations of this sort clearly in mind does not provide a cure for the limitations, but it does reduce the prospect of misuse, misinterpretation, or exaggeration. For brevity, we refer to this as the issue of transparency.

Data

The data used in the case study are the product of the Montréal Longitudinal Study of Boys (Tremblay, Desmarais-Gervais, Gagnon, & Charlebois, 1987). The 1,037 male subjects in this study were in kindergarten at the study’s outset in the spring of 1984.

They were next assessed in 1988 and then again annually until 1995 when their average age was 17. The sample was drawn from 53 schools in the lowest socioeconomic areas in Montréal, Canada. To control for cultural effects, the longitudinal study included boys only if both their biological parents were born in Canada and their biological parents’ mother tongue was French. This resulted in a homogeneous White, French-speaking sample. Wide-ranging measurements of potentially important covariates, such as social and psychological function, were based on assessments by parents, teachers, and peers; self-reports of the boys themselves; and administrative records from schools and the juvenile court. These measurements include data on the boys’ behavior across many domains (e.g., sexual activity and delinquency in adolescence) and social functioning (e.g., peer popularity). (See Tremblay et al., 1987, for further details on this study.)

The self-reported data on annual involvement in violent delinquency and participation in delinquent groups, which we hereafter refer to as gangs, form the core of the analyses we report in this article. Queries on prior year involvement in violent delinquency and gangs were initiated in 1989 when the boys were age 11. Subjects were asked about the frequency of their involvement in seven different types of violent delinquency within the past year—threatening to attack someone, fist fighting, attacking an innocent person, gang fighting, throwing objects at people, carrying weapons, and using weapons. These items were each coded on a 4-point Likert scale (0 = *never*, 1 = *once or twice*, 2 = *sometimes*, 3 = *often*) and summed to form an overall scale of violent delinquency. This scale was used to estimate the trajectory model from ages 11 to 13. In estimating the treatment effect of gang membership, the item pertaining to gang fighting was excluded. Gang membership status in the prior year was based on the subject’s response to the question, “During the past 12 months, were you part of a group or a gang that committed reprehensible acts?”¹

A total of 580 individuals in the Montréal study reported no involvement with gangs from ages 11 through 13 and also had no more than one missing assessment of their violent delinquency and gang involvement over this period.² These individuals form the basis for the analyses reported here.

Group-Based Trajectory Modeling

The group-based trajectory model is an application of finite mixture modeling. As described in Land, McCall, and Nagin (1996), Nagin (2005), Nagin and Land (1993), and Nagin and Tremblay (2001, 2005), the method is designed to approximate the population distribution of developmental trajectories with a finite number of trajectory groups. In our case study the trajectory groups are approximating the distribution of violent delinquency trajectories from ages 11 to 13 of boys with no history of gang membership. Each trajectory group is meant to characterize a salient feature of this distribution and thereby create an interesting

¹ The original version of the question as administered in French was: “Au cours des 12 derniers mois, as-tu fais partie d’un groupe de jeunes [gang] qui fait des mauvais coups?”

² Prior to age fourteen, 282 boys were involved in gangs, and 128 had more than one missing assessment over this age range. An additional 59 boys were missing either their gang membership status at age 14 or their assessment of violent delinquency.

Table 1
Three-Group Trajectory Model

Variable	Group 1	Group 2	Group 3
	Low	Declining	Chronic
Group probability (π_j)	.463	.478	.060
Expected rate of violence at 11 (λ_{11}^j)	0.309	1.88	4.66
Expected rate of violence at 12 (λ_{12}^j)	0.281	1.53	4.35
Expected rate of violence at 13 (λ_{13}^j)	0.255	1.24	4.05
% in gangs at 14	7.0 (21 of 297)	14.7 (38 of 254)	31.0 (9 of 29)

Note. Low = low violence; Declining = declining violence; Chronic = chronic violence.

and useful stratification of pretreatment levels of the outcome variable.

The rationale for approximation is elaborated in the citations in the previous paragraph. The essence of the argument is that social science theory rarely provides much guidance on the exact form of the population distribution of trajectories. One use of finite mixture models is to approximate a continuous distribution function whose exact form is unknown (Heckman & Singer, 1984; McLachlan & Peel, 2001; Titterton, Smith, & Makov, 1985). Pickles and Angold (2003) noted that

both theoretical and empirical work in statistics has shown that the nonparametric estimator of the underlying distribution, essentially the best fitting distribution, is just such a set of discrete classes of this kind, *even when the underlying distribution is continuous*. (p. 541; emphasis in original)

For this reason, the group-based trajectory method is often described as a semiparametric method (e.g., Nagin, 2005; Nagin & Tremblay, 1999, 2001).

For readers who are unfamiliar with group-based trajectory, we refer them to Muthén (2001, 2004) or Nagin (1999, 2005) for technical elaboration. For our purposes here it suffices to describe three key outputs of the model: (a) the form of each group's trajectory, which is usually defined by a polynomial function of age whose order may vary across group; (b) the size of each group as measured by the proportion of the population following that trajectory; and (c) the posterior probability of group membership in each trajectory group j . For each individual i in the sample, (c) is the probability that that individual's measurement history (e.g., violent delinquency from ages 11 to 13) was the product of trajectory group j . We denote the proportion of the population in each trajectory group by π_j , where j is the index of trajectory group, and the posterior probabilities by π_{ij} , where i is the index of each sampled individual.

Stage 1: Creating Trajectory Groups to Measure the Trend Before Treatment

For our illustrative application, the trajectory model describes violence from ages 11 to 13. The underlying model requires the specification of the distribution within trajectory group and by age of this outcome variable. We assume that within each trajectory group j , self-reported violent delinquency at each age is Poisson distributed. We further assume that the logarithm of the expectation of this Poisson variable, λ^j , follows a linear function of age for each trajectory group. Model estimation requires specification of

the number of trajectory groups. For this application the number of groups was set at three.³

Two details of the model deserve mention. First, as noted, the model was estimated under the assumption that within trajectory group the violence delinquency scale varied according to a Poisson process. Although the scale is not, strictly speaking, a count variable, it resembles a Poisson random variable in several respects. The violence score takes values in the nonnegative integers, is rightward skewed, and—most importantly for our purposes—within trajectory group, the means and variances of the scale at each age are about equal, as is true of the Poisson distribution. Second, the model is estimated without random effects. As just noted, the equality of the mean and variance within groups suggests there is no further within-group variation. Also, the addition of random effects to a group-based model can result in the use of fewer trajectory groups, because their addition allows for more within-group heterogeneity. Because the trajectory groups are intended to define clusters of individuals following approximately the same developmental course, this potential increase in within-group heterogeneity is counterproductive.

Table 1 reports the estimates of π_j and λ^j from the three-group model. Group 1, which we call the low-violence group, was estimated to make up 46.3% of the population. The estimates of λ^j for this group declined only very slightly with age, and the slope coefficient estimate underlying this decline was not statistically significant at the .10 level. The second group, which we call the

³ Because just three periods of data were used to estimate the trajectory model, only linear trajectory models were estimated. We used the Bayesian information criterion to select the number of trajectory groups in the model. The Bayesian information criterion values for the three- and four-group models were about equal. We used the three-group model for two reasons. The first was parsimony; the second concerned the instability of the four-group model. The four-group model split one group in the three-group model without materially altering the other two groups. Haviland and Nagin (2005) found that when the model was repeatedly estimated across bootstrapped samples, the variability of the parameter estimates of the three-group model was consistent with the sampling variability implied by the maximum likelihood estimates of their standard errors. By contrast, the parameter estimates for the two groups carved from what is called the declining group in the three-group model were very unstable across the bootstrap samples. The parameters of this model are estimated by a direct maximum likelihood procedure available in SAS (Jones & Nagin, in press; Jones, Nagin, & Roeder, 2001). The procedure accommodates missing assessments under the assumption that they are missing completely at random.

declining group, followed a trajectory in which the expected rate of violence declined with age. For this group the slope coefficient estimate was significant at the .01 level. This group was estimated to make up 47.8% of the population. The final group was composed of a small contingent of individuals estimated at 6.0% of the population whose rate of violence was high and nearly constant. We call this the chronic group.

Table 1 also reports a cross tabulation of trajectory group membership with other measurements on the sample members.⁴ Observe that there is a pronounced association of trajectory group membership from ages 11 to 13 with gang membership at age 14. This association exemplifies the self-selection problem that our approach is designed to overcome.

Stage 2: Estimating the Propensity Score and Checking the Balance in Measured Covariates After Matching

In the opening paragraphs of this article we characterized estimation of the violence facilitation effect of gang membership as a prototypical example of a much larger class of problems that are central to the study of human development—inferring the effect of a turning-point event or therapeutic intervention on the developmental course of an outcome of interest. In the parlance of experimental design, one of the main challenges of making causal inferences with observational data about the effect of treatment on outcome is that both may be affected by preexisting psychosocial characteristics of the individual. In our analysis the treated are first-time gang members at age 14 and the controls are individuals who continue to refrain from gang membership at 14. Figure 1 provides concrete grounding into the inference problem. It reports box plots comparing the treated and controls on covariates measured at baseline—age 13 or earlier. The box plots make clear that most are systematically related to gang membership status at age 14. Before joining gangs, joiners were more violent than nonjoiners; were less popular with their peers; were more aggressive, hyperactive, and oppositional; had more sexual partners; and had mothers whose age at the birth of their first child was younger. If this had been a randomized experiment in which the boys in the Montréal study had been selected at random to join a gang at age 14, one covariate in 20 would be expected to be significantly related to gang membership status for $\alpha = .05$. Ten of the 12 covariates in Figure 1 have significance levels less than $\alpha = .05$.

Each of these characteristics and behaviors is also systematically related to violence and physical aggression (Lacourse, Nagin, Tremblay, Vitaro, & Claes, 2003; Lacourse et al., 2006; Nagin & Tremblay, 2001; Thornberry et al., 2003; Tremblay & Nagin, 2005). Thus, if these and other characteristics are not properly accounted for in the analysis, they may bias the effect of treatment on outcome.

As previously described, propensity score matching is designed to achieve covariate balance between treated and controls on all covariates included in the score. There are two key theorems concerning propensity scores demonstrated in Rosenbaum and Rubin (1983). Informally, they say (a) matching or stratifying on the unidimensional propensity score tends to balance the observed covariates used to construct the score, and (b) if there is no bias from unobserved covariates, then adjusting for the propensity score is sufficient to obtain a treatment effect free of confounding. The propensity score is thus a device for removing imbalances in

observed pretreatment covariates. Unlike randomization, matching on a propensity score does little or nothing to control imbalances from covariates that were not measured, and these remain a central concern. Through matching or stratification on the propensity score, one compares treated and control groups that appear comparable prior to treatment in terms of observed covariates. Our strategy is to match gang joiners with comparable gang nonjoiners using propensity scores where the matches are restricted to occur within trajectory group.

As a first step in matching the gang joiners with comparable nonjoiners, we estimated a propensity score using the original 12 covariates in Figure 1 plus some others derived from these covariates; see Haviland et al. (2007) for details about the preliminary propensity score in this figure. (We describe in detail, below, the estimation of the propensity actually used in the matching reported here.)

Figure 2 reports box plots of preliminary estimated propensity scores for joiners and potential controls in each of the three trajectory groups. In the low and declining trajectory groups, there is a substantial difference between joiners and potential controls, but there is also a fair amount of overlap in the distributions of joiners and potential controls. This implies good matches should be available, which indeed is the case. By contrast, in the chronic group, the distributions exhibit very limited overlap. The lower quartile among the joiners, for example, was above the upper quartile among the nonjoiners.

The results in Figure 2 suggest that it would be difficult to find good matches for the gang joiners in the chronic group. This indeed turned out to be the case. Haviland et al. (2007) found that all attempts to find good matches for the gang joiners in the chronic group failed dismally. They concluded that these data do not permit credible estimation of the effect of gang membership for the individuals in the chronic trajectory. Thus, we restrict our analysis to the more numerous but less violent individuals in the low and declining groups. The final propensity score model was estimated using only the members of the low and declining trajectory groups.

Before turning to a discussion of the matching process for these two trajectory groups, two points deserve commentary. First, one could try to skirt the problem of the limited supply of matches for gang joiners in the chronic group by running a regression to adjust for the covariate difference between joiners and nonjoiners in the chronic group and hoping for the best. However, for this group the regression would largely consist of an extrapolation between joiners and nonjoiners whose covariate distributions exhibit limited overlap. By definition, extrapolation of this sort compares individuals who were not comparable prior to treatment. A key limitation of regression unaided by matching is that it does not warn the user when it is extrapolating. For example, regression permits the comparison of the health status of a treated group under 20 years old and a control group over 65 years old, “adjusting for age,” even though such a comparison is preposterous.

Second, the stratification into trajectory groups provided a natural definition of the subpopulations in which comparable boys can be compared. Although investigators sometimes define the

⁴ Individuals were assigned to the trajectory group for which their posterior probability of group membership was largest.

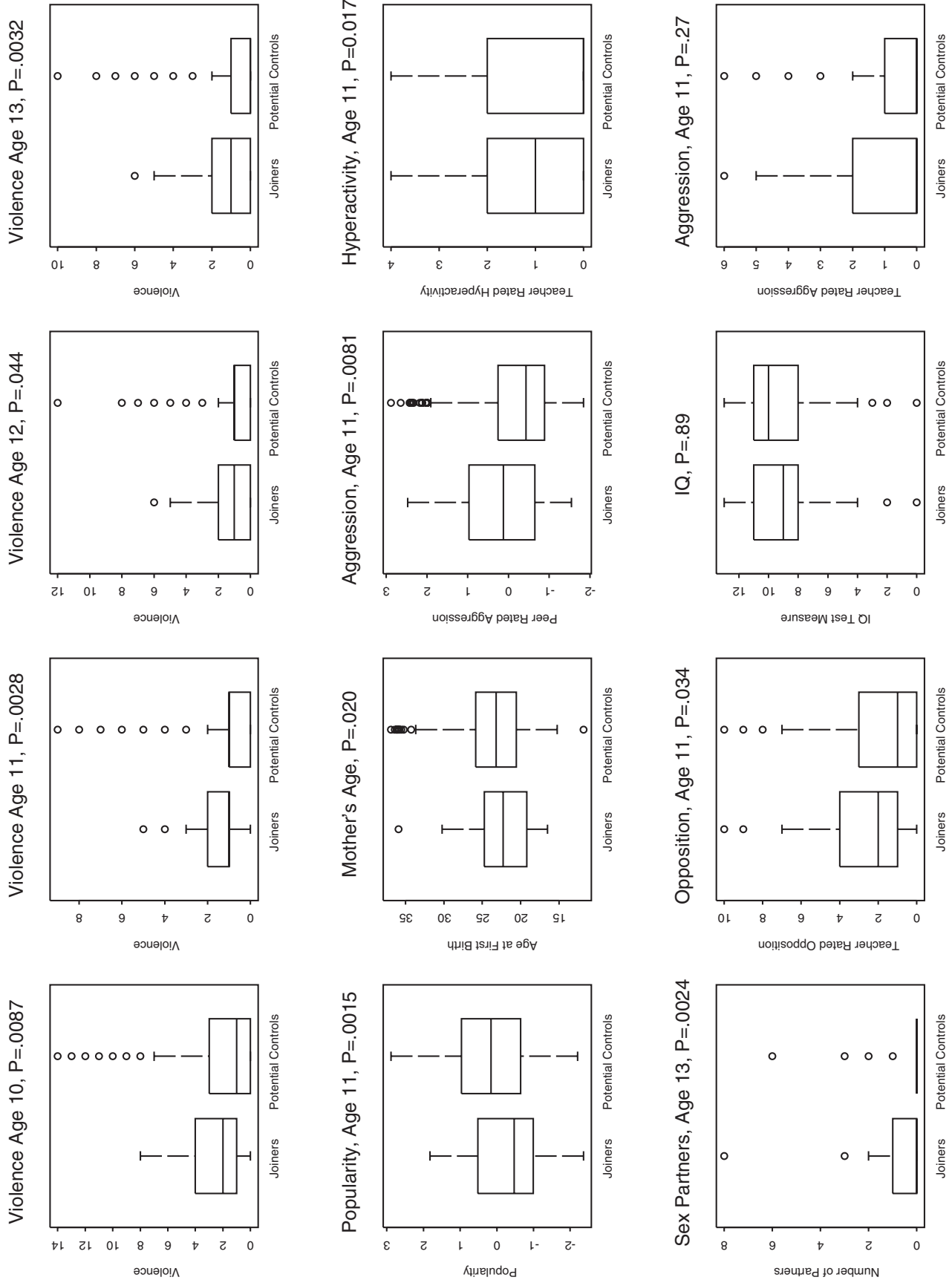


Figure 1. Box plots of 12 covariates before matching for gang joiners at age 14 and for potential controls who did not join at age 14. The p value is from Wilcoxon's two-sided rank sum test. Because of discreteness, two quartiles are sometimes equal.

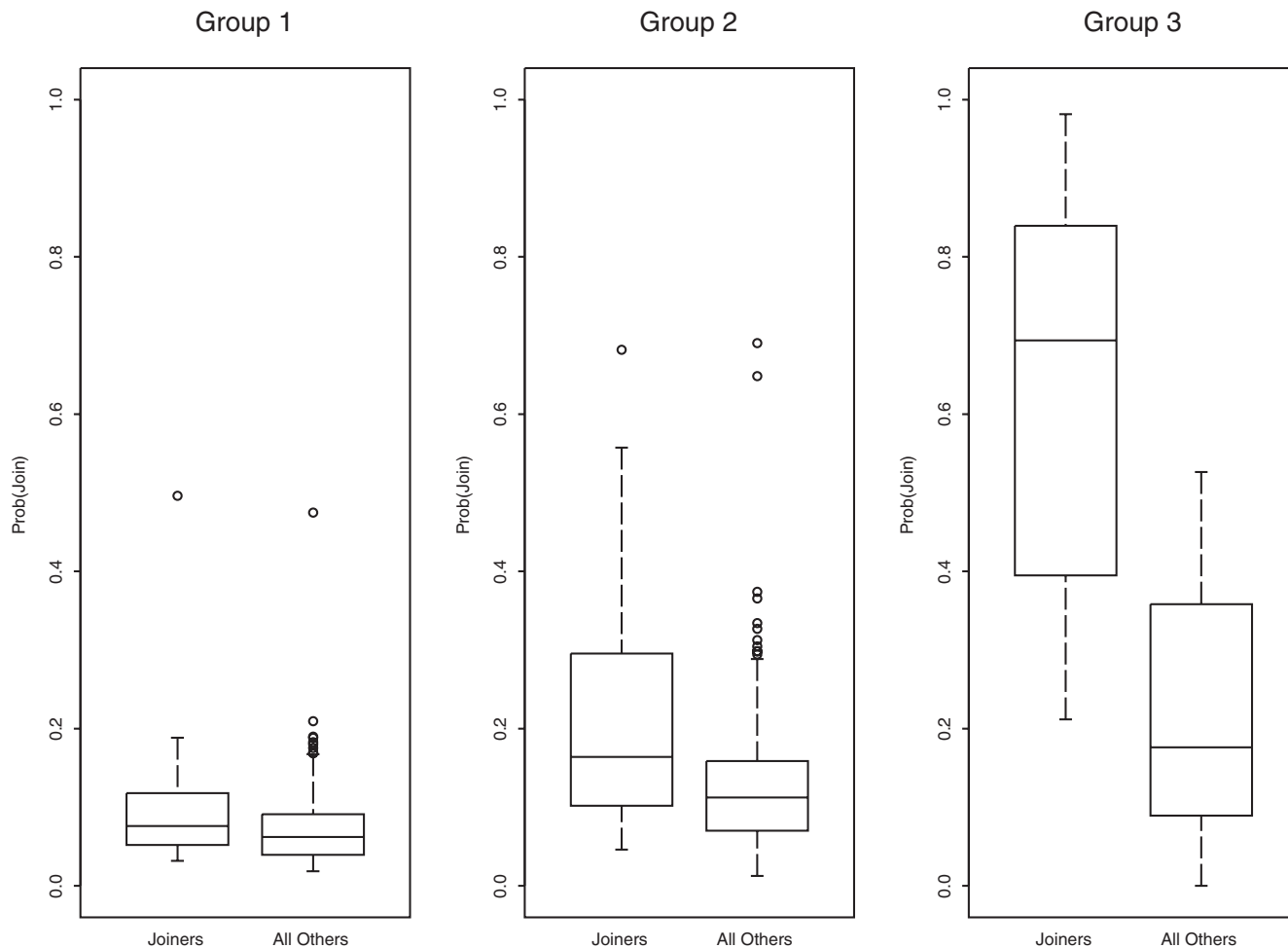


Figure 2. Box plots of the first estimated propensity scores by trajectory group. Low is trajectory group $s = 1$, Declining is group $s = 2$, and Chronic is group $s = 3$. In later analyses, the propensity score was reestimated without group $s = 3$. prob = probability.

region of overlap on covariates using the propensity score, such a region may lack a clear substantive interpretation; in contrast, trajectory groups are defined by the outcome under study, and so they often form a highly interpretable way to define the region of overlap.

Matching on Propensity Score and Selected Covariates

Once obtained, the propensity scores of treated and controls can be matched in different ways. In this analysis, we matched each gang joiner with 2 nonjoiners who had approximately the same propensity score and similar values on other key covariates. Matching is done without replacement so that a nonjoiner is not matched to more than 1 joiner. We briefly discuss two issues concerning this match strategy: (a) the mechanics of the matching algorithm and (b) the choice of the number of matches.

For each gang joiner a multivariate distance between that individual and each potential control was computed based on the propensity score and 10 of the most important covariates in the score.⁵ The final distance between a joiner and a potential control

was this multivariate distance, plus a large penalty if these 2 individuals had propensity scores differing by more than a quarter of a standard deviation of the propensity score, plus a second large penalty if they also differed by more than half of a standard deviation of the propensity score. Optimal matching assigned 2 controls to every joiner, with no control used more than once, to minimize the total distance within matched triples. The matching was performed with the OPTMATCH package in the statistical software package R (Hansen, 2007; R Development Core Team, 2007).

As shown in Haviland et al. (2007), matching with 2 controls rather than pair matching substantially improves precision. In fact, matching with 2 controls improves precision over pair matching by half the distance to matching with infinitely many controls. Match-

⁵ The distance is the Mahalanobis distance, invented by the Indian statistician P. C. Mahalanobis, and it is the multivariate analog of the square of a standardized difference. The multivariate version uses covariances as well as variances in the standardized difference.

ing with large numbers of controls, say more than 10 controls, does little to further improve efficiency, and it becomes much harder to find acceptable matches. For a full discussion of these issues, see Smith (1997) and Ming and Rosenbaum (2000). Ming and Rosenbaum in addition address the benefits of matching variable numbers of controls across treated units, a matching strategy that is demonstrated in Haviland et al. (2007).

Results of the Match

Table 2 reports statistics on balance before and after matching for the covariates included in the propensity score. The standardized biases pre- and postmatching are computed as follows:

$$d = \frac{|\bar{x}_t - \bar{x}_c|}{s_x},$$

$$d^m = \frac{|\bar{x}_t - \bar{x}_c^m|}{s_x},$$

$$s_x = \sqrt{s_{xt}^2 + s_{xc}^2/2},$$

where \bar{x}_t , \bar{x}_c , and \bar{x}_c^m are the means of covariate x for, respectively, all treated units, all potential controls, and all controls included in the match, and s_{xt}^2 and s_{xc}^2 are the sample variances of x for all treated and all potential control units, respectively. The quantities d and d^m , respectively, denote the standardized bias for the unmatched and matched samples.

For d the numerator is the absolute difference between the sample average of x for the treated units and that for all potential controls. For d^m the numerator is the absolute difference between the sample average of x for the treated units and that for their matched controls. The standardized bias measures the size of these two differences relative to a measure of the variability of x in the treated and in the pool of potential control units. It is calculated as the square root of the average of s_{xt}^2 and s_{xc}^2 . Ideally, the standardized bias is zero. In a randomized experiment, the treated-minus-control difference in means of a covariate has expectation zero, but in each actual experiment, d will differ somewhat from zero due to the luck of the randomization. In a randomized experiment, when compared using a two-sample test such as the t test, only 1 in 20 differences in covariate means is expected to be significant at the $p < .05$ level, in marked contrast to what typically happens in an observational study (see Figure 1). The aim of matching in an observational study is to reduce the standardized bias.

Before matching, the treated and control groups were about two thirds of a standard deviation apart on the propensity score, almost half a standard deviation apart on peer-rated popularity, almost 40% of a standard deviation apart on the posterior probability of membership in the declining trajectory group, and roughly one quarter of a standard deviation apart on violence at ages 10, 11, 12, and 13. Twelve of the covariates in Table 2 had absolute standardized differences d strictly greater than 0.2 or 20% of a standard deviation before matching, but none of the d^m were strictly greater

Table 2
Covariate Imbalance Before and After Matching

Covariate	d	d^m	t	t^m	\bar{x}_t	\bar{x}_c	\bar{x}_c^m
Logit propensity score	0.64	0.17	3.96	0.95	0.17	0.10	0.15
Pr (Low Low or Declining)	0.38	0.06	2.80	-0.41	0.65	0.50	0.68
Total violence ^a	0.34	0.07	2.37	0.42	5.56	4.28	5.31
Violence							
Age 10	0.24	0.18	1.76	1.20	2.46	1.96	2.07
Age 11	0.24	0.02	1.79	0.11	1.23	0.93	1.20
Age 12	0.23	0.19	1.48	-1.07	0.99	0.73	1.21
Age 13	0.23	0.07	1.61	0.39	0.89	0.67	0.83
Peer-rated popularity, age 11	0.46	0.09	-3.50	-0.64	-0.23	0.18	-0.15
Peer-rated aggression, age 11	0.25	0.14	1.78	0.85	-0.01	-0.23	-0.13
Teacher rating of hyperactivity, age 11	0.21	0.19	1.54	1.18	1.22	0.95	0.98
Teacher rating of opposition, age 11	0.19	0.16	1.35	1.04	2.49	2.03	2.10
Teacher rating of physical aggression, age 11	0.03	0.17	-0.20	1.17	0.76	0.79	0.55
Mother's age at first birth	0.25	0.02	-1.90	0.12	22.59	23.56	22.52
Number of sexual partners, age 13	0.20	0.16	1.39	0.96	0.23	0.14	0.16
Intelligence score	0.07	0.09	-0.50	-0.56	8.92	9.10	9.14
Missing data							
Intelligence score	0.16	0.12	0.98	0.64	0.03	0.01	0.02
Mother's age	0.13	0.06	0.81	0.30	0.03	0.01	0.03
Aggression rating	0.09	0.02	0.62	0.15	0.15	0.12	0.14
Popularity	0.09	0.02	0.62	0.15	0.15	0.12	0.14
Physical aggression	0.15	0.00	0.96	0.00	0.05	0.02	0.05
Number of sexual partners	0.27	0.20	1.54	1.09	0.05	0.01	0.02
Violence age 13	0.14	0.11	0.89	0.64	0.03	0.01	0.02
Mean	0.23	0.10					
Maximum	0.64	0.20					

Note. d = standardized bias for the unmatched sample; d^m = standardized bias for the matched sample; t^m = t statistic after matching; \bar{x}_t = the mean of covariate x for all treated units; \bar{x}_c = the mean of covariate x for all potential controls; \bar{x}_c^m = the mean of covariate x for all matched controls.

^a The total violence variable is the sum of the self-reported violence scores at ages 10 to 13. This variable is among those included in the multivariate distance measure but is not in the propensity score model because the violence measures at each age are in the model.

than 0.2 after matching, so the matched groups were much more comparable than the original unmatched groups.

Also reported in Table 2 are conventional two-sample t -test statistics for a test of the hypothesis that the means are different before and after matching, using data from the low and declining trajectory groups. As noted, in a completely randomized experiment, 1 in 20 such t statistics on covariates is expected to yield a significant difference at the $p < .05$ level, so the two-sample t statistics compare the balance on covariate means to the balance in a randomized experiment of the same size. For instance, in Table 2, none of the 22 t statistics t^m after matching is 2 in absolute value, so the balance on these covariates is somewhat better than expected in a completely randomized experiment. Of course, the data in Table 2 are not from a completely randomized experiment; that is just a yardstick for comparison. Because perfect balance on covariate means is not available even in a completely randomized experiment, perfect balance is not the appropriate yardstick; rather, a completely randomized experiment is one plausible yardstick. Mistakes when using these t statistics are common and are critically important to avoid. Both t before matching and t^m after matching compare the observed imbalances in means to the anticipated chance imbalances in corresponding completely randomized experiments with the corresponding sample sizes. From Table 2, the imbalance before matching was much worse than expected in a completely randomized experiment with the given sample

size: There were significant differences in the propensity score, in the probability of being in the low group, in total violence, and in peer-rated popularity. The imbalance in means after matching was somewhat better than expected in a completely randomized experiment with the sample size of the 2-to-1 matched study. It is a mistake, however, to compare t to t^m , because the sample sizes are different. For instance, in Table 2, matching did little to improve the balance on teacher rating of hyperactivity at age 11, as $d = 0.21$ while $d^m = 0.19$ are both about 20% of a standard deviation, but $t = 1.54$ while $t^m = 1.18$ is somewhat reduced, largely because the sample size has fallen. Moreover, it is a mistake to see the t statistic as a formal test of balance when applied to estimated quantities, such as the propensity score, which were estimated from the data to discriminate treated and control groups. The t statistics are an informal yet useful yardstick, nothing more.

Figure 3 depicts the covariates themselves for the four covariates in Table 2 with the largest standardized biases before matching. Figure 3 compares the 59 boys who joined gangs, their $2 \times 59 = 118$ matched controls, and the remaining unmatched boys. The gang joiners and their matched controls look similar, and both look quite different from the unmatched boys. In short, the matching seems to have worked in removing imbalances in the distributions of the four variables with the largest imbalances before matching.

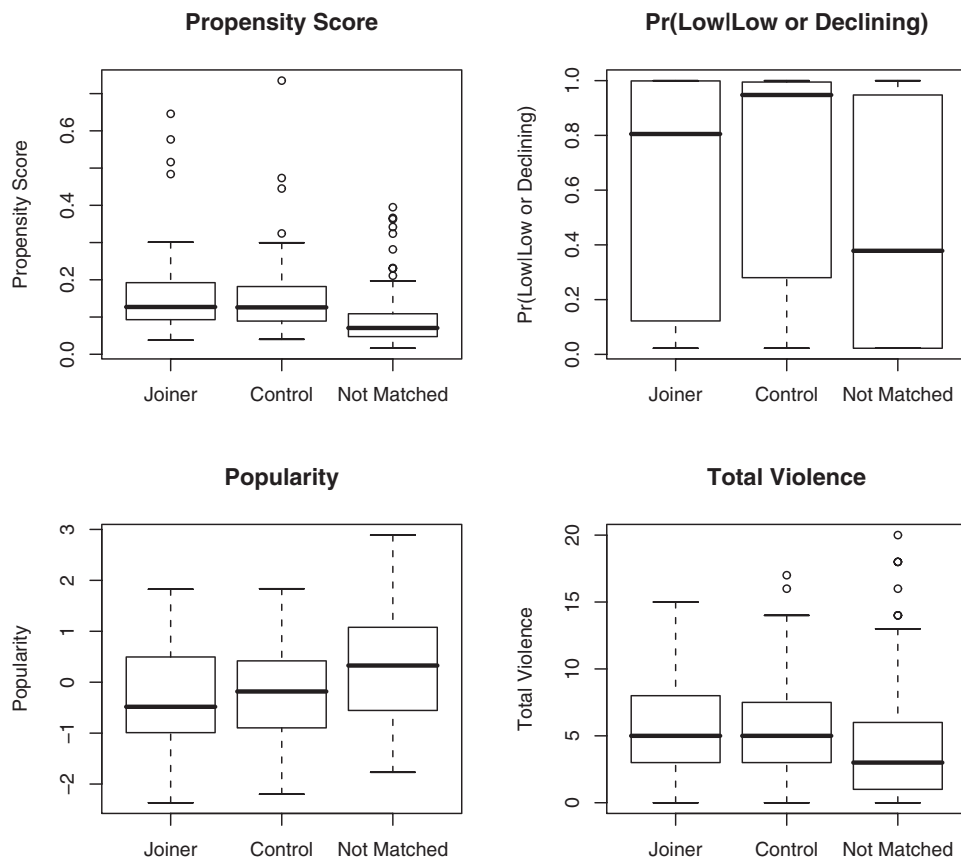


Figure 3. Box plot of four covariates with large biases before matching, comparing the gang joiners, their matched controls, and the remaining unmatched boys in the low and declining trajectory groups.

Stage 3: Analyzing the Effect of Gang Membership at Age 14 on Violence at Age 14 and Beyond

We turn now to examining the effect of first-time gang membership at age 14 on violence at age 14 and beyond. In the analytic strategy demonstrated here, matching within trajectory group is used to control for potentially confounding measured covariates. Results in Stage 2 suggest that the matched treated and control boys were similar before treatment with respect to the covariates that were measured. In particular, they were similar on prior measures and trajectories of the outcome variable, delinquent violence, which is key among the measured covariates as a potential confounder.

After successful matching, an unadjusted contrast of the violence of first-time gang members at age 14 and their matched abstainers provides an estimate of the gang facilitation effect at age 14 and beyond. Let y_i^t denote the outcome for the i th treated participant (or gang joiner) and y_{ik}^c the outcome for the k th control matched to joiner i . In our analysis we matched 2 controls to each treated case, so $k = 1, 2$. For each treated participant the estimate of the treatment effect is simply the difference between y_i^t and the average of i 's 2 matched controls, \bar{y}_i^c . For each trajectory group the estimated treatment effect is simply the average of these differences across treated cases assigned to the group:

$$\bar{\Delta} = \frac{1}{N} \sum_i (y_i^t - \bar{y}_i^c), \tag{1}$$

where N is the number of treated cases in a trajectory group.⁶

Table 3 reports estimates of the facilitation effect for ages 14 to 17. The effect is calculated according to Equation 1; namely, it is the average of $y_i^t - \bar{y}_i^c$. Results are reported for the low and declining trajectory groups combined and separately. They show that at age 14 the gang joiners had markedly higher rates of violence than the

Table 3
The Effects of Gang Membership at Age 14 on Violence From Ages 14 to 17

Age (in years) and group	Violence difference (Δ)	One-sided p value
14		
Low	0.81	.012
Declining	0.91	.019
Combined	0.87	.002
15		
Low	0.60	.30
Declining	0.76	.057
Combined	0.69	.037
16		
Low	0.57	.075
Declining	-0.39	.370
Combined	-0.05	.863
17		
Low	0.61	.175
Declining	0.02	.962
Combined	0.24	.402

Note. Low = low violence; Declining = declining violence; Chronic = chronic violence.

abstainers for both trajectory groups and for their combination. Figures 4 and 5 report box plots comparing the distributions of violence at ages 14 to 17 of gang joiners and their matched control abstainers. The box plots for age 14 show a clear upward shift of the violence distribution of the joiners compared to the abstainers for both trajectory groups. After age 14, however, the effects of gang membership at age 14 seem to dissipate. At age 15, the age 14 joiners continued to have significantly higher rates of violence (one-tailed test for $\alpha = .05$). However, for the individual trajectory groups the violence differences at age 15 were not significant. At ages 16 and 17 there were no significant differences.

In an experiment, the integrity of a treatment is the extent to which participants assigned to that treatment continue in that treatment as intended by the study protocol. In many longitudinal experiments, participants fail to comply with the full course of treatment, eroding the integrity of the treatment group. The nature and effects of many longitudinal treatments are affected by integrity. A treatment may be ineffective because participants typically drop out, a less potent treatment may prove more effective because participants stick with it, and so on. We briefly explore the extent to which erosions of treatment integrity are present in this observational longitudinal data by exploring the following questions: What is the integrity of joining a gang at age 14? Is it a major and permanent transition or a brief and transient event? To what extent is treatment integrity related to the violence pattern in Figure 4?

Table 4 provides some answers. The table reports gang membership status at ages 14, 15, 16, and 17 by membership status at age 14. The results indicate that, at least in Montréal, gang membership as measured in this study is highly transient. At age 15, fewer than 40% of the first-time joiners at age 14 remained in a gang. By age 17 only about 20% reported being in gang. Also, a sizable minority of the age-14 abstainers subsequently became gang members—14% at age 15, 17% at age 16, and 13% at age 17. Boys who joined gangs at age 14 soon quit; others soon joined. Given this, the dissipation of effect with time in Figure 4 is not surprising. If a boy were prevented from joining a gang at age 14, not much would be changed: There is an excellent chance he would have quit soon if he had joined, and also an excellent chance that he would join at later age anyway.

This example illustrates what we believe is one particularly valuable feature of the approach we have demonstrated—keeping time in order. In an experiment, both intended outcomes—violence at age 14 and beyond and also persistence in treatment (gang membership at age 15 and beyond)—are outcomes of initial treatment status (gang membership at age 14). A key finding of our analysis was that gang membership was highly transient. In contrast, some methods for analyzing longitudinal data would view gang membership throughout time as an external or ancillary time-varying treatment and violence throughout time as an outcome. If adjustments are made for an outcome of treatment as if it were a covariate measured before treatment, then estimates of treatment effects are often biased; see Rosenbaum (1984). As is seen from Robins, Greenland, and Hu (1999) and the ensuing discussion, issues of this sort can creep into longitudinal analyses that intend to avoid them. The method we used here and the more

⁶ To minimize notation, we have not subscripted quantities by trajectory group.

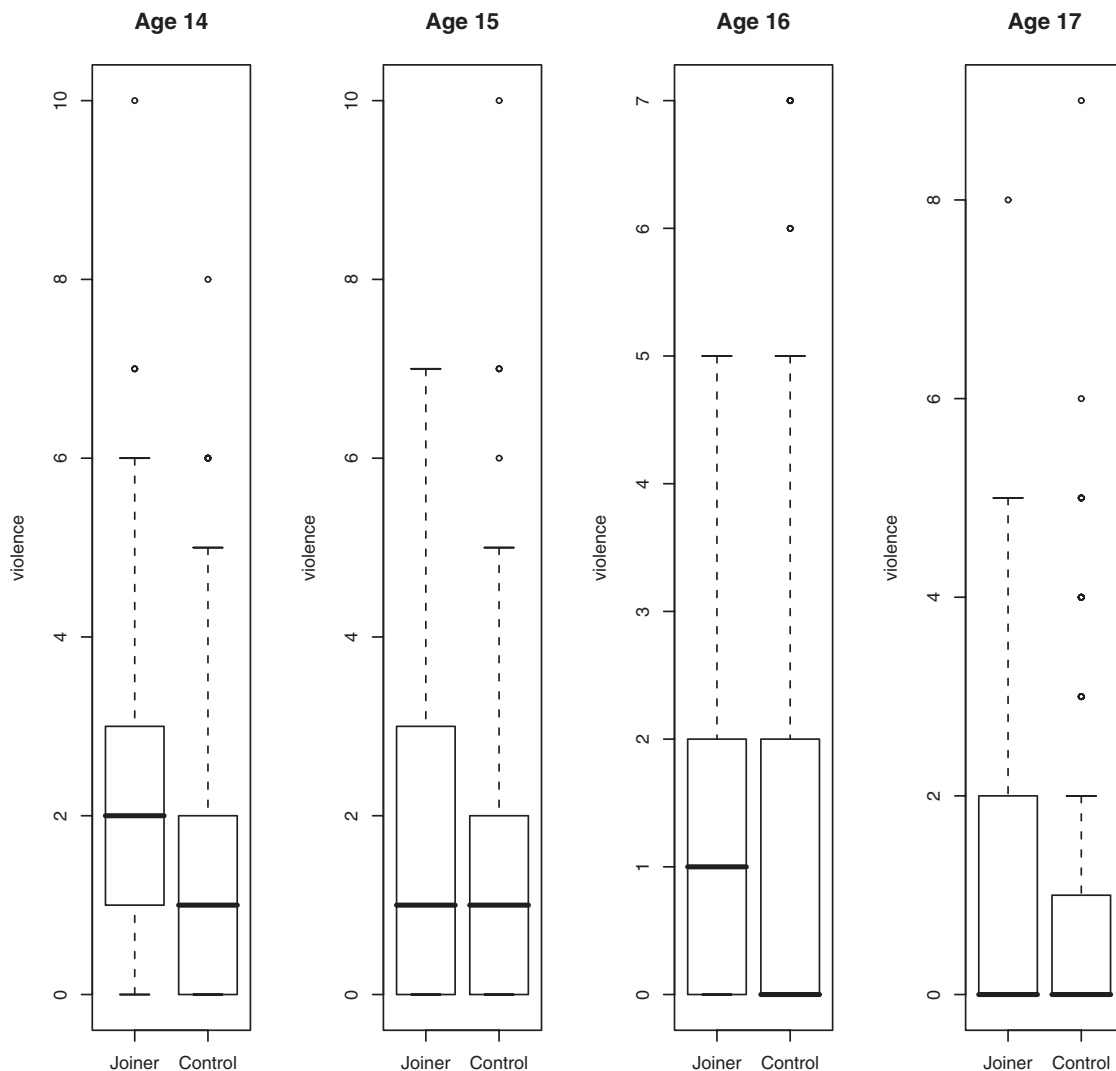


Figure 4. Violence at ages 14 to 17 for boys who joined gangs at age 14 and their matched controls who did not join gangs at age 14. Low and declining trajectory groups are combined.

general strategy of risk-set matching (Li et al., 2001; Lu, 2005) both stress a transparent experimental analog, in which covariates are prior to the start of treatment, outcomes are subsequent to the start of treatment, and persistence or not in treatment is an outcome, not a covariate or external time-varying treatment.

Haviland et al. (2007) performed three other related analyses. They demonstrated an approach to calibrating the effect of treatment as received that uses treatment assigned as an instrumental variable in a treatment integrity or noncompliance analysis; see also Greevy, Silber, Cnaan, and Rosenbaum (2004). Haviland et al. also considered potential biases from covariates that were not measured. They asked what such a covariate would have to be like to alter the conclusions of the study—that is, they performed a sensitivity analysis. They found the increase in violence at age 14 for joiners is insensitive to small unobserved biases but is sensitive to biases of moderate size. See Rosenbaum (1991, 2002, 2005c) for discussion of sensitivity analysis, and see Rosenbaum (2005a, 2005b, 2006) for discussion of methods for reducing sensitivity to

bias from unobserved covariates. Finally, Haviland et al. combined matching with robust covariance adjustment for observed covariates, with a view to increased efficiency.

Conclusions

In this article we have demonstrated an approach to making causal inferences based on observational data about the effect of a turning-point event on the developmental trajectory of the behavior under study. Making causal inferences with nonexperimental data is fraught with ambiguity in almost all problem contexts. In developmental studies much of this ambiguity takes a particular form, namely, the prior trajectory of the behavior under study predicts both the likelihood of the turning-point event and the future trajectory of the behavior. The approach described here is designed to address this form of the problem by trying to take full advantage of the rich set of measurements that are a hallmark of modern longitudinal studies. Group-based trajectory modeling is

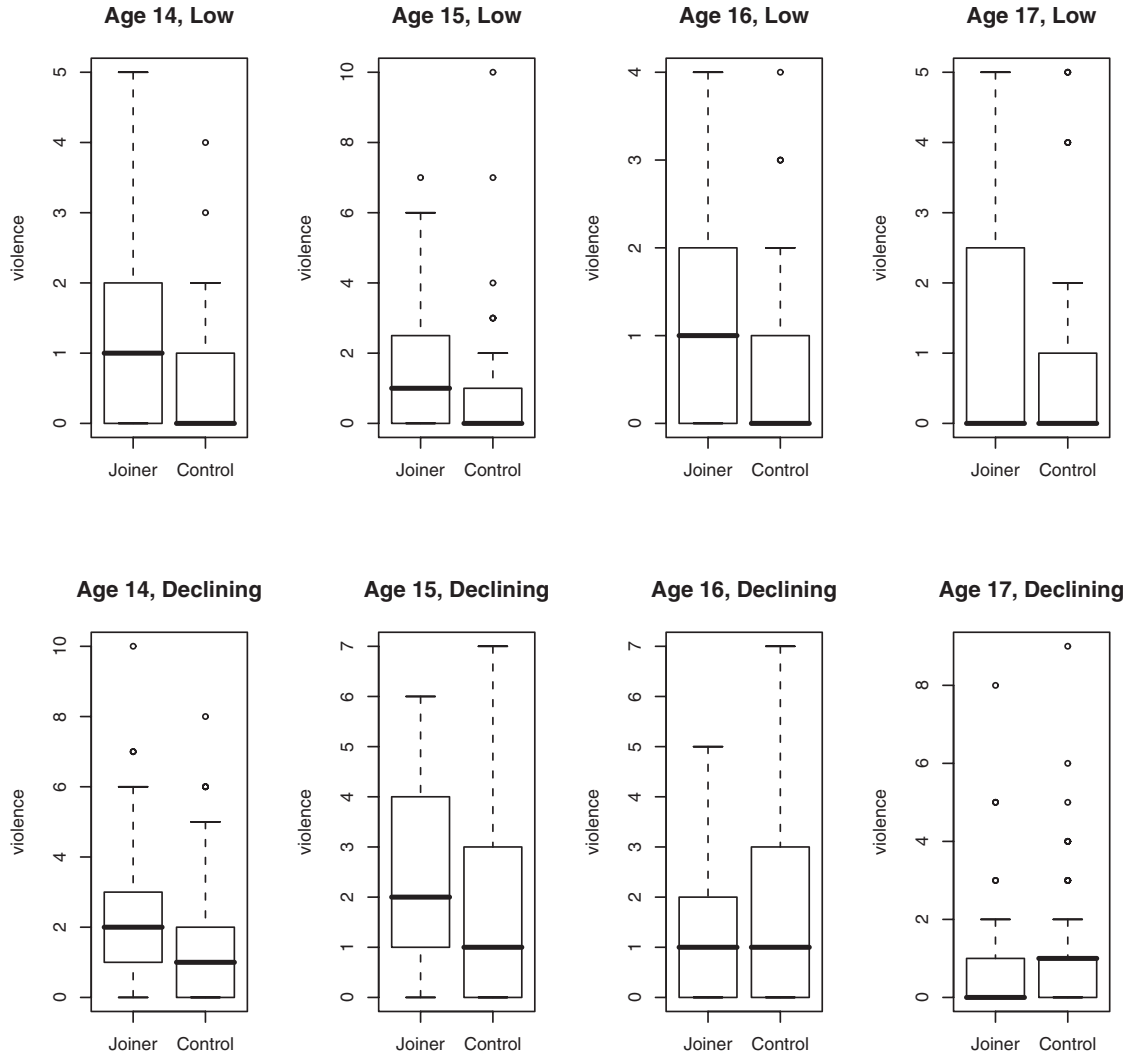


Figure 5. Violence at ages 14 to 17 for boys who joined gangs at age 14 and their matched controls who did not join gangs at age 14. Trajectory groups are separated.

used to stratify the data into clusters of individuals with similar trajectories of behavior prior to their experiencing the turning-point event. Stratification by trajectory group also provides a means for identifying developmentally meaningful subgroups in the population for whom treatment effects may vary. Propensity score matching is used to insure balance on the large numbers of

measured covariates present in these data that might otherwise bias the treatment effect estimate.

The approach is also designed to bring some of the key attributes of an experiment to the design and analysis of non-experimental data. Experiments, we believe, have especially high standing in adjudicating the validity of scientific claims for two reasons. One is well understood: If properly conducted, an experiment is not vulnerable to the argument that estimated effects are contaminated by biases from confounding influences. The other is less commonly appreciated—it is transparency. Cox (1958) observed that in an experiment “simplicity . . . is a very important matter . . . which must be constantly born in mind” (p. 11). We agree. Experimental results can be presented in very simple forms—for example, a comparison of the mean outcomes of alternative treatment conditions. This makes the results accessible even to individuals with no statistical training.

Table 4
Gang Membership Status at Ages 15, 16, and 17 for Joiners at Age 14 and Their Matched Controls

Age 14 gang status	% in gangs by age (in years)			
	14	15	16	17
Gang member	100	38.6	25.5	21.2
Nonmember	0	14.2	16.5	13.3

Although trajectory groups, propensity scores, and optimal matching have some technical details, the end product is simply presented and understood—treated and control groups that are comparable at baseline, prior to treatment, with respect to observed covariates. Elementary methods, such as box plots and means, suffice to check whether this comparability on observable covariates has been achieved. Similarly, outcomes in treated and control groups, overall and within trajectory group, may be compared and tested with elementary statistical methods.

The simplicity of the presentation also has the salutary effect of keeping the key weakness open to public view and discussion. For example, we determined that we could not find similar matches for the small number of treated subjects in the chronic violence trajectory, and thus our results are only applicable to youth exhibiting lower levels of violence. In addition, although our treated and control groups are quite comparable with respect to the covariates in Table 2, they may well differ in terms of covariates that were not measured. Table 2 is a transparent and carefully circumscribed assertion of comparability: The boys were comparable in these respects, but perhaps not in others. The ease with which the results can be critically scrutinized without resort to technical tools or terminology is a strength, not a weakness, of the approach we have demonstrated. It serves to broaden the audience that can participate in the critical assessment of the findings and the range of tools that may be used in that critical assessment. For instance, in an observational study of mortality after surgery in the Medicare population, Rosenbaum and Silber (2001) suggested combining quantitative analysis of many matched pairs with qualitative or ethnographic analysis of a small number of pairs, with a view to understanding comparability in ways that are not part of the available quantitative data.

In brief, the method we proposed has the following strengths. It compares treated and control subjects who appeared comparable at baseline before treatment with respect to measured covariates, including developmental trajectory prior to treatment, and it provides an internal check that this type of comparability has been achieved. This strength is particularly compelling when the rich set of measurements present in many modern longitudinal datasets is available. The method provides substantively meaningful subgroups within which overlap in the covariate distributions can be checked and groups in which there is not adequate overlap can be identified. Differential treatment effects can also be estimated within these substantively meaningful subgroups. Like a longitudinal experiment, the method keeps time in order, with the integrity of treatment explicitly an outcome of treatment—often an important outcome as the effects of a treatment typically depend upon its integrity. The method is transparent, with its achievements and weakness equally open to public view and discussion; specifically, the inevitable possibility of bias from a covariate that was not measured is front and center for discussion.

References

- Cox, D. R. (1958). *Planning of experiments*. New York: Wiley.
- Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.
- Elder, G. H., Jr. (1985). Perspectives on the life course. In G. H. Elder, Jr. (Ed.), *Life course dynamics* (pp. 23–49). Ithaca, NY: Cornell University Press.
- Elder, G. H., Jr. (1998). The life course as developmental theory. *Child Development*, *69*, 1–12.
- Gail, M. H. (1972). Does cardiac transplantation prolong life? A reassessment. *Annals of Internal Medicine*, *76*, 815–817.
- Greevy, R., Silber, J. H., Cnaan, A., & Rosenbaum, P. R. (2004). Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial. *Journal of the American Statistical Association*, *99*, 7–15.
- Hansen, Ben B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News*, *7*(2), 19–24.
- Haviland, A., & Nagin, D. S. (2005). Causal inference with group-based trajectory models. *Psychometrika*, *70*, 1–22.
- Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory modeling in an observational study. *Psychological Methods*, *12*, 247–267.
- Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, *52*, 271–320.
- Joffe, M. M., & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology*, *150*, 327–333.
- Jones, B. L., & Nagin, D. S. (2007). Advances in group-based trajectory modeling and a SAS procedure for estimating them. *Sociological Methods and Research*, *35*, 542–572.
- Jones, B., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Research and Methods*, *29*, 374–393.
- Lacourse, E., Nagin, D., Tremblay, R. E., Vitaro, F., & Claes, M. (2003). Developmental trajectories of boys' delinquent group membership and facilitation of violent behaviors during adolescence. *Development and Psychopathology*, *15*, 183–197.
- Lacourse, E., Nagin, D. S., Vitaro, F., Côté, S., Arseneault, L., & Tremblay, R. E. (2006). Prediction of early onset deviant peer group affiliation: A 12-year longitudinal study. *Archives of General Psychiatry*, *63*, 562–568.
- Land, K., McCall, P., & Nagin, D. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models with empirical applications to criminal careers data. *Sociological Methods and Research*, *24*, 387–439.
- Li, Y. P., Propert, K. J., & Rosenbaum, P. R. (2001). Balanced risk set matching. *Journal of the American Statistical Association*, *96*, 870–882.
- Lu, B. (2005). Propensity scores with time dependent covariates. *Biometrics*, *61*, 721–728.
- Maguin, E., Loeber, R., & LeMahieu, P. G. (1993). Does the relationship between poor reading and delinquency hold for males of different ages and ethnic groups? *Journal of Emotional and Behavioral Disorders*, *1*, 88–100.
- McLachlan, G. J., & Peel, D. (2001). *Finite mixture models*. New York: Wiley.
- Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, *56*, 118–124.
- Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent curve modeling. In A. Sayer & L. Collins (Eds.), *New methods for the analysis of change* (pp. 291–322). Washington, DC: American Psychological Association.
- Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Sage handbook of quantitative methodology* (pp. 345–368). Thousand Oaks, CA: Sage.
- Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods*, *4*, 139–177.
- Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

- Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, *31*, 327–362.
- Nagin, D. S., Pagan, L. S., Tremblay, R. E., & Vitaro, F. (2003). Life course turning points: The effect of grade retention on physical aggression. *Development and Psychopathology*, *15*, 343–361.
- Nagin, D. S., & Tremblay, R. E. (1999). Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. *Child Development*, *70*, 1181–1196.
- Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods*, *6*(1), 18–34.
- Nagin, D. S., & Tremblay, R. E. (2005). Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology*, *43*, 873–904.
- Nieuwebeerta, P., Nagin, D. S., & Blokland, A. (2007). *The relationship between first imprisonment and criminal career development: A matched sample comparison*. Unpublished manuscript.
- Pagan, L., Boulerice, B., Vitaro, F., & Tremblay, R. E. (1999). Effects of poverty on academic failure and delinquency in boys: A change and process model approach. *Journal of Child Psychology and Psychiatry*, *40*, 1209–1219.
- Pickles, A., & Angold, A. (2003). Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Development and Psychopathology*, *15*, 529–551.
- R Development Core Team. (2007). R: A language and environment for statistical computing [Computer software]. Vienna: R Foundation for Statistical Computing.
- Robins, J. M., Greenland, S., & Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association*, *94*, 687–712.
- Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A*, *147*, 656–666.
- Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine*, *115*, 901–905.
- Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.
- Rosenbaum, P. R. (2005a). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician*, *59*, 147–152.
- Rosenbaum, P. R. (2005b). Observational study. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1451–1462). New York: Wiley.
- Rosenbaum, P. R. (2005c). Sensitivity analysis in observational studies. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1809–1814). New York: Wiley.
- Rosenbaum, P. R. (2006). Differential effects and generic biases in observational studies. *Biometrika*, *93*, 573–586.
- Rosenbaum, P. R., Ross, R. N., & Silber, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, *102*, 75–83.
- Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
- Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.
- Rosenbaum, P., & Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*, 33–38.
- Rosenbaum, P. R., & Silber, J. H. (2001). Matching and thick description in an observational study of mortality after surgery. *Biostatistics*, *2*, 217–232.
- Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, *27*, 325–353.
- Thornberry, T., Krohn, M., Lizotte, A., Smith, C., & Tobin, K. (2003). *Gangs and delinquency in developmental perspective*. Cambridge, United Kingdom: Cambridge University Press.
- Titterton, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.
- Tremblay, R. E., Desmarais-Gervais, L., Gagnon, C., & Charlebois, P. (1987). The preschool behavior questionnaire: Stability of its factor structure between culture, sexes, ages, and socioeconomic classes. *International Journal of Behavioral Development*, *10*, 467–484.
- Tremblay, R. E., and Nagin, D. S. (2005). Aggression in humans. In R. E. Tremblay, W. W. Hartup, & J. Archer (Eds.), *Developmental origins of aggression* (pp. 83–106). New York: Guilford.

Received April 27, 2007

Revision received September 24, 2007

Accepted October 2, 2007 ■